Détection et visualisation des communautés dans les réseaux sociaux augmentés

Juan David Cruz Gomez, Cécile Bothorel, François Poulet

Détection et visualisation des communautés dans les réseaux sociaux

Juan David Cruz[1], Cécile Bothorel[1], François Poulet[2]

1. Département LUSSI, Télécom-Bretagne
Technopôle Brest-Iroise, 29238, Plouzané, France
juan.cruzgomez@telecom-bretagne.eu

2. Université de Rennes 1, IRISA
Campus de Beaulieu, 35042, Rennes, France
francois.poulet@irisa.fr

RÉSUMÉ. Les réseaux sociaux contiennent trois types de variables : structurelles, représentant les interactions entre acteurs, de composition ou sémantiques, qui décrivent les caractéristiques de chaque acteur, et d'affiliation, qui sont utilisées pour représenter l'appartenance des nœuds aux différents groupes. À l'heure actuelle, la plupart des approches d'analyse des réseaux sociaux n'utilisent que l'information structurelle et d'affiliation, en perdant les descriptions sémantiques de chaque acteur. Cet article propose un modèle d'intégration de l'information provenant d'un réseau social, qui permet d'identifier et de visualiser les communautés avec des critères sémantiques et structurels.

ABSTRACT. Social networks contain, or can be described by, three types of variables: structural variables, which represent information about the relationships between all actors taken as a whole, composition variables, which contain characterizing information about actors as individuals, and affiliation variables, describing nodes according to their belongings to different groups. Most existing approaches use only structural and affiliation variables, leaving behind semantic descriptions of the actors of the network. This paper proposes a model to integrate all information composing a social network, which allows identifying and visualizing communities with semantic and structural criteria.

MOTS-CLÉS : détection de communautés, visualisation de communautés, analyse de réseaux sociaux.

KEYWORDS: community detection, community visualization, social network analysis.

DOI:10.3166/RIA.26.369-392

1. Introduction

Un réseau social est composé d'un groupe d'acteurs liés par différents types de relations. Toutefois, l'information contenue dans un réseau social n'est pas limitée aux seuls nœuds et arêtes : elle inclut aussi de l'information sémantique, utilisée pour décrire chaque acteur de manière individuelle.

Ainsi, les méthodes d'analyse de réseaux sociaux utilisent en général chaque type d'information séparément, mais, grâce à l'intégration des différentes variables composant les réseaux sociaux, il est possible d'identifier de nouvelles informations qui n'étaient pas visibles avant.

Nous proposons une nouvelle approche pour intégrer les informations présentes dans un réseau, puis, cette intégration réalisée, un modèle de détection et de visualisation de communautés qui permet d'établir des relations entre l'information sémantique et structurelle. Ce modèle visuel permet de mettre en évidence les interactions entre les communautés et guide l'identification des nœuds ayant un rôle important pour la communication. Les rôles proposés sont adaptés de Cross et Parker (2004) :

– nœuds centraux : avec un haut degré de connectivité, et par lesquels passe la plupart de l'information ;

– boundary spanners : font office de pont entre les nœuds de différentes communautés ;

– information brokers : peuvent être vus comme les experts d'un certain domaine ;

– peripheral people : personnes qui n'interagissent pas avec des personnes des autres communautés (ou qui n'ont aucun lien avec d'autres individus).

Ces rôles apparaissent visuellement en fonction de la position du nœud dans le réseau. En intégrant les informations du graphe il est également possible d'identifier les nœuds qui changent de rôle d'une configuration à l'autre.

Cet article est organisé de la façon suivante : la section 2 présente les précédents travaux connexes, dans la section 3 le modèle d'intégration de l'information contenue dans un réseau social est proposé, dans la section 4 une analyse de la complexité du modèle est présentée. Dans la section 5 quelques expériences sont présentées avant la conclusion et les travaux futurs.

2. Travaux précédents

Les réseaux sociaux sont composés de trois types de variables (Wasserman, 1994) :

1. Variable structurelle : variable qui représente le graphe et les connexions entre les nœuds. Les mesures associées sont en général calculées sur tout le graphe.

2. Variable de composition : variable utilisée pour décrire chaque nœud du réseau à partir de ses attributs, par exemple, sexe, nationalité, niveau d'éducation.

3. Variable d'affiliation : variable destinée à décrire l'appartenance de chaque nœud à un des groupes identifiés dans le réseau.

En général, l'analyse des réseaux sociaux est dédiée à la mesure de chacune de ces variables et à l'identification des nœuds importants ou des relations non triviales.

La détection de communautés constitue un type d'analyse mettant en œuvre les trois variables décrites précédemment, la troisième variable étant générée à partir de la structure, de la variable de composition ou des deux en même temps.

2.1. Détection de communautés

La plupart des travaux en détection de communautés sont basés sur la structure du graphe : l'idée de base est de trouver des partitions dont le nombre d'arêtes à l'intérieur des groupes est supérieure au nombre d'arêtes en dehors des groupes. Par conséquent, chaque groupe ainsi identifié possède une forte connectivité à l'intérieur et une faible connectivité avec les autres groupes.

Cette idée a été résumée par Gaetler (2005) qui a défini trois indices pour mesurer la qualité des partitions trouvées : la *couverture*, qui mesure le poids de toutes les arêtes dans les groupes *versus* le poids des arêtes connectant les groupes ; la *conductance*, fondée sur l'observation que si un groupe est bien connecté, alors un nombre important d'arêtes doivent être enlevées pour le diviser en deux parties égales ; enfin, la *performance* indique si les deux extrémités d'une arête appartiennent au même groupe ou si deux nœuds ne formant pas une arête appartiennent à des groupes différents (le groupement est alors dit correct). Une autre mesure, utilisée de plus en plus souvent (Fortunato, 2010) est la *modularité*. Cette mesure, proposée par Newman et Girvan (2004), mesure la proportion des arêtes dans les groupes *versus* le nombre des arêtes en dehors des groupes, elle compare aussi cette proportion avec celle d'une partition aléatoire du même graphe.

La méthode présentée par Newman et Girvan (2004) trouve et enlève, de façon itérative, l'arête avec le degré d'intermédiation le plus élevé. Ce processus permet de trouver des groupes connectés de façon souple avec d'autres groupes mais fortement connectés à l'intérieur de chacun d'eux. La partition sélectionnée est celle avec la modularité la plus haute ; toutefois le calcul direct de la modularité a une complexité en $O(n^2)$, où n est le nombre de nœuds du graphe, ce qui fait que le temps de calcul pour l'algorithme est important.

Blondel *et al.,* (2008) proposent une approche pour la détection de communautés définie comme un problème d'optimisation dont la fonction objectif est la maximisation de la modularité. La méthode commence par affecter chaque nœud à une communauté « singleton » en générant une première partition et en calculant la valeur initiale de modularité. Puis, chaque nœud est changé de communauté vers une autre communauté voisine, et pour tout changement la nouvelle modularité est calculée : si la nouvelle valeur est positive et supérieure à la valeur précédente, le nœud est affecté à la communauté qui maximise la modularité. Si aucun changement n'améliore la modularité, le nœud reste dans sa communauté d'origine. Ce processus est effectué

jusqu'à ce qu'aucun changement ne soit plus possible. Le coût de cet algorithme est linéaire par rapport au nombre d'arêtes dans le cas de graphes non denses.

Le travail proposé par Du *et al.,* (2007) s'intéresse à la détection de communautés dans les réseaux sociaux à grande échelle. La méthode commence par énumérer toutes les cliques maximales (des graphes complets qui ne sont contenus dans aucun autre sous-graphe complet) ; après l'énumération des cliques, elle génère des noyaux qui font office de centre de gravité où les nœuds sont affectés selon leur proximité avec chacun de ces noyaux. La complexité est en $O(n^2)$, où n est le nombre de nœuds du graphe.

La plupart des méthodes trouvent des partitions disjointes, toutefois dans les applications des réseaux sociaux il est tout à fait possible de trouver des nœuds qui appartiennent simultanément à plusieurs communautés. Pizzuti (2009) présente un algorithme génétique dont le but est de maximiser le nombre d'arêtes dans chaque groupe. Lipczak et Milios, (2009) présentent aussi un algorithme génétique pour identifier des communautés floues. La différence fondamentale est liée à la définition de la fonction d'adaptation, qui est composée par une combinaison de trois mesures : le *normalized cut* par Shi et Malik (2000), la largeur de *silhouette* (Rousseeuw, 1987) et la *modularité*. La complexité de ces méthodes est en $O(n^2 + \lambda)$, où λ est le nombre d'itérations des simulations.

Les méthodes précédentes n'utilisent que les variables structurelles du réseau social. Néanmoins les communautés peuvent être aussi identifiées à partir des variables de composition du réseau. Le travail de Zhou *et al.,* (2009) présente une méthode pour utiliser simultanément les variables structurelles et de composition pour identifier les communautés. Leur méthode utilise un surfeur aléatoire qui traverse k groupes prédéfinis et qui essaye de maximiser la distance entre les groupes en affectant les nœuds à chaque groupe selon leur similarité. Pour ce faire un graphe augmenté est créé, qui représente les variables de composition du graphe. Avec la matrice de probabilité de transition du graphe augmenté, le surfeur aléatoire trouve les k groupes sémantiquement proches. La qualité structurelle est mesurée par la densité d'arêtes dans chaque groupe.

2.2. Visualisation de graphes de communautés

Un graphe de communautés est un type de graphe hiérarchique dont la distance entre deux nœuds de l'arbre d'inclusion est au maximum égale à un.

La Figure 1 présente un exemple de graphe de communautés et la Figure 2 présente l'arbre d'inclusion de la partition. Cet arbre est utilisé pour visualiser les différents niveaux qui représentent les groupes.

Cette définition est proposée et utilisée par Eades et Feng (1997) lors de la présentation de leur méthode de visualisation. Cette méthode représente chaque niveau de l'arbre d'inclusion comme une couche tridimensionelle. D'abord, pour la première couche, les nœuds feuilles de chaque communauté sont placés ; dans la deuxième couche, un cercle est ensuite dessiné au-dessus de chaque groupe. Ce

processus est répété jusqu'à la racine de l'arbre d'inclusion. La figure 3 présente un exemple de la visualisation avec la méthode de Eades et Feng.

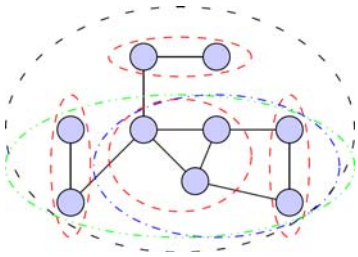

Figure 1. Exemple de graphe de communautés

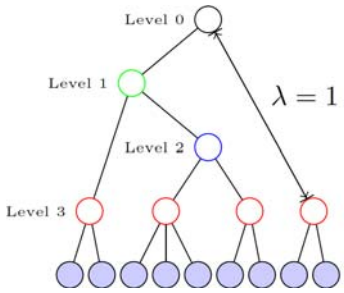

Figure 2. Arbre d'inclusion de la hiérarchie de groupes

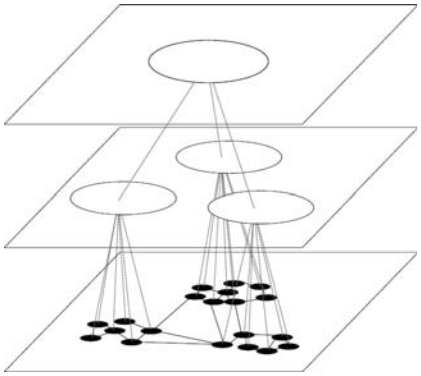

Figure 3. Exemple de graphe de communautés dessiné avec l'algorithme de Eades and Feng

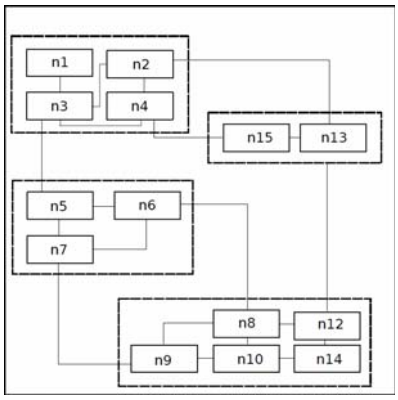

Figure 4. Exemple d'un graphe de communautés dessiné avec la représentation orthogonale

Tamassia (1987) présente un algorithme qui utilise des rectangles pour représenter les nœuds et des lignes droites avec des angles droits pour représenter les arêtes ; l'auteur utilise cette approche pour dessiner des circuits VLSI. Les groupes sont représentés par des rectangles autour des nœuds. Un exemple de cette représentation est montré dans la figure 4. L'algorithme présenté par di Giacomo *et al.,* (2007) utilise aussi une approche orthogonale, mais appliquée à la représentation de sites web.

Noack, (2003) présente un algorithme dirigé par des forces qui utilise d'une part un modèle linéaire pour représenter l'attraction entre deux nœuds voisins et, d'autre part, un modèle logarithmique pour modéliser la répulsion entre deux nœuds non

connectés. Ce modèle n'utilise pas un graphe déjà groupé : il constitue aussi une méthode pour identifier graphiquement des communautés dans un réseau social.

Bourqui *et al.* (2007) présentent une méthode de visualisation qui prend en compte les poids des arêtes. Cette méthode impose quatre contraintes :

– interdire le chevauchement des groupes pour faciliter l'interprétation du dessin,

– garder l'arbre d'inclusion pour visualiser la hiérarchie produite par l'algorithme de détection de communautés,

– utiliser un polygone convexe pour définir chaque groupe,

– respecter les poids du graphe en utilisant des fonctions de minimisation de l'énergie.

Avec la définition de graphe de communautés de Eades et Feng (1997) et avec la définition de graphe quotient de Brockenauer et Cornelsen (2001), les auteurs commencent par placer les nœuds individuels, puis les nœuds des niveaux suivants. À l'inclusion de chaque niveau, l'espace de visualisation est divisé en utilisant des diagrammes de Voronoï. La Figure 5 présente un exemple de ce modèle de visualisation.

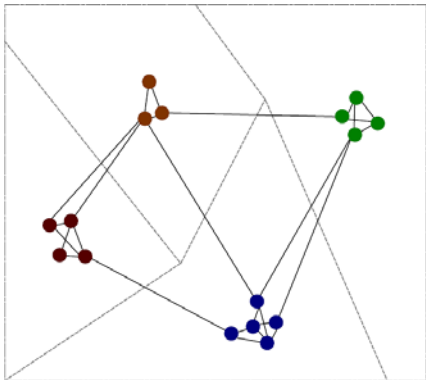

Figure 5. Exemple de visualisation du graphe pondéré de communautés

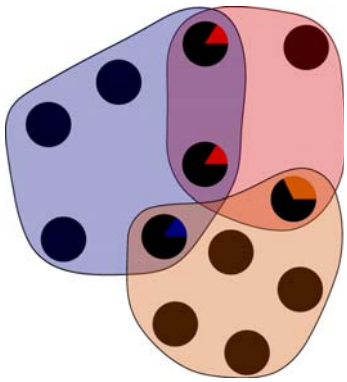

Figure 6. Exemple de graphe de communautés chevauchantes

En général les méthodes de visualisation de graphes de communautés ont été conçues pour mettre en évidence les différences entre chaque groupe, ce qui implique de présenter chaque groupe éloigné des autres. Santamaría et Therón (2008) proposent une méthode pour dessiner des communautés non disjointes. Ils utilisent une version modifiée de l'algorithme dirigé par des forces dont la force d'attraction est modélisée par un ressort et la répulsion par un modèle de gravitation. Pour éviter l'encombrement des éléments visuels, les arêtes ne sont pas dessinées, par contre les groupes sont modélisés par des polygones convexes. Ainsi, les nœuds appartenant à différents groupes sont placés à la frontière de ces polygones de telle façon que le partage soit évident. La Figure 6 présente un exemple de cette

visualisation ; les nœuds partagés sont dessinés comme un camembert indiquant le degré d'appartenance à chaque groupe.

Les algorithmes classiques n'utilisent que la structure du graphe pour placer les nœuds et en général il n'est pas facile de leur ajouter d'autres critères, comme la similarité entre nœuds, pour trouver la position de chacun des nœuds.

2.3. Définition du problème

Dans les deux sections précédentes nous avons présenté les principaux travaux en détection et visualisation de communautés dans les réseaux sociaux. Dans le premier cas la plupart des algorithmes n'utilisent que la structure du graphe pour identifier les communautés, et, d'un autre côté, les méthodes de visualisation de communautés ont été conçues pour accentuer les frontières entre chaque groupe.

Il existe alors une déconnexion des variables contenues dans le réseau social ; l'intégration de ces variables va permettre d'identifier et d'extraire de nouvelles connaissances à partir de l'analyse du réseau.

Nous cherchons à répondre à la question : comment intégrer et exploiter les variables des réseaux sociaux ?

Nous proposons donc une méthode d'intégration des variables structurelle et de composition pour produire une nouvelle variable d'affiliation. En plus de l'intégration des variables, nous proposons une méthode de visualisation qui mette en évidence les relations et interactions entre les communautés du réseau.

3. Description du modèle

Le modèle proposé permet d'intégrer les variables structurelles et de composition du graphe, puis de visualiser les communautés en soulignant les interactions entre ces communautés. La figure 7 présente le modèle général du système.

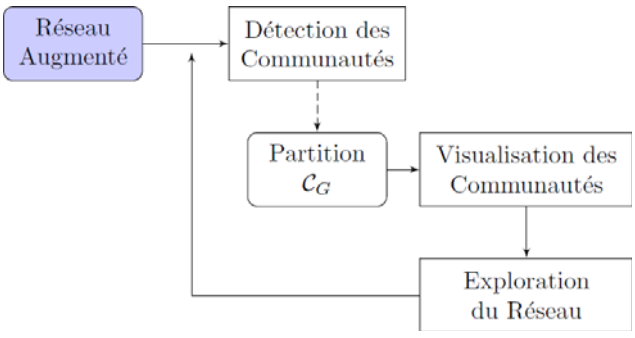

Figure 7. Diagramme général du modèle

L'entrée du système est un réseau augmenté, composé des variables structurelles et d'un ensemble de caractéristiques pour chaque nœud. Le processus de détection de communautés produit une partition du graphe qui intègre les deux variables précédentes. Cette partition est ensuite utilisée par l'algorithme de visualisation pour dessiner les communautés et permettre l'exploration.

3.1. Quelques définitions

Un réseau augmenté G(V, E, F*) est représenté comme un graphe non orienté, avec un ensemble de nœuds V, un ensemble d'arêtes E et $F^* \in \mathbb{R}^f$ l'ensemble de caractéristiques de la variable de composition (f étant sa taille). Soit $C=\{C_1, C_2, \ldots, C_k\}$ une partition de V, $e(u,v) \in E$ est l'arête entre les nœuds *u* et *v*, e(u) est l'ensemble des arêtes du nœud *u*. Soit $E(C_i, C_j)$, $1 \leq i,j \leq k$, l'ensemble des arêtes entre les communautés *i* et *j*. $E(C_i, C_i) = E(C_i)$, $1 \leq i \leq k$ est l'ensemble des arêtes de la communauté *i*, $E(C_i, C_i) \subset E$.

Étant donné que les variables de composition F* peuvent être composées de $f > 0$ éléments, le fait d'avoir un très grand ensemble de caractéristiques peut mener à des résultats inattendus à cause de la réduction de la signification statistique par l'effet Hughes ou malédiction de la dimension (Hughes, 1968), ainsi que par l'addition de bruit dérivé de l'utilisation de caractéristiques non compatibles. Pour réduire ces effets nous proposons le concept de *point de vue*.

Définition 3.1 (Point de vue POV_F). Étant donné un ensemble de caractéristiques F, un point de vue POV_F est une des 2^f combinaisons des f éléments de F*.*

Ainsi, le point de vue est un sous-ensemble de F*. En conséquence, le graphe peut être décrit depuis différentes perspectives, chacune étant définie par un point de vue.

Définition 3.2 (Nœud intérieur v^-). Étant donné un groupe $C_i \in C$, un nœud v est dit intérieur à C_i si et seulement si $\forall \varepsilon \in e(v), \varepsilon \in E(C_i)$.

Cela signifie que le nœud intérieur a uniquement des liens avec d'autres nœuds de sa propre communauté. La figure 8 présente un exemple de nœuds intérieurs.

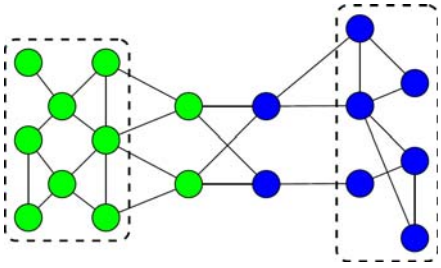

Figure 8. Exemple de nœuds intérieurs

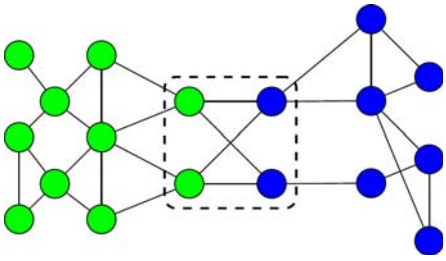

Figure 9. Exemple de nœuds de frontière

Définition 3.3 (Nœud de frontière v^+). Étant donné un groupe $C_i \in C$, un nœud v est dit de frontière de C_i si et seulement si $\exists \varepsilon \in e(v), \varepsilon \notin E(C_i)$.

Ainsi les nœuds de frontière ont des liens vers/depuis d'autres communautés. La figure 9 présente un exemple de nœuds de frontière.

3.2. Détection de communautés

L'intégration des variables structurelles et de composition permet de guider le processus d'identification des communautés en ajoutant la similarité de chaque nœud aux critères de la structure utilisés pendant l'identification des communautés. Pour ce faire, nous divisons le processus de détection de communautés en deux phases : d'abord, un clustering de nœuds selon leur similarité sémantique, puis, un algorithme de clustering structurel exploitant la partition sémantique, c'est-à-dire influencé de façon telle que les groupes contiennent les nœuds similaires et connectés.

3.2.1. Première phase : clustering sémantique

Étant donné un point de vue dérivé d'un ensemble POV_F, chaque nœud peut être caractérisé par son vecteur d'attributs ou une instance *u* du point de vue. Il est possible d'utiliser ces vecteurs en entrée d'un algorithme de classification non supervisée comme les cartes auto-organisatrices (Kohonen, 1997). Cela permet de créer des groupes de nœuds suivant la similarité de leurs attributs, c'est-à-dire les instances de *u* sont les données en entrée de l'algorithme. L'avantage de cet algorithme (conseillé par Newman et Girvan (2004) pour les algorithmes de détection de communautés) est que, à la différence des approches comme les k-means, l'utilisateur n'a pas besoin de fixer a priori le nombre final de groupes.

L'algorithme de cartes de Kohonen utilisé a un réseau N basé sur une grille rectangulaire de taille de f x f neurones, avec $f = | POV_{F^*} |$, le nombre d'attributs utilisés dans le point de vue. Les valeurs initiales des poids sont tirées aléatoirement. Les poids des neurones sont ajustés selon leur proximité au neurone gagnant. Un taux d'apprentissage η est utilisé pour éviter les maxima locaux et des convergences prématurées. Après chaque itération, le taux d'apprentissage est réduit par un facteur ε, $0 < \varepsilon < 1$. Le voisinage est calculé avec une taille *t* et le neurone gagnant est de centre *c*.

La sortie est alors une partition C_{SOM} formée par des groupes de nœuds similaires en termes du point de vue choisi. Pour mesurer la qualité de la partition nous utilisons la distance moyenne entre les points de chaque groupe, laquelle a été mise à l'échelle pour avoir des valeurs entre 0 et 1.

3.2.2. Deuxième phase : influence et clustering structurel

Une fois que la partition sémantique C_{SOM} a été calculée, on peut alors entrer dans la seconde phase de la méthode. Dans cette étape, on utilise un algorithme classique de détection de communautés, le *fast unfolding algorithm – FU*, proposé

par Blondel *et al.* (2008) et présenté dans la section 2. Cet algorithme utilise la modularité *Q*, présentée par Newman et Girvan (2004) comme mesure de qualité.

Avant l'exécution du *fast unfolding algorithm,* on inclut les informations obtenues lors de la première phase. Cela est effectué par le changement des poids des arêtes en fonction de la partition obtenue C_{SOM}. Pour chaque paire de sommets v_i, $v_j \in V$, $\forall\ v_i \neq v_j$, le poids de l'arête $e(v_i \neq v_j)$ est modifié par la distance euclidienne des instances du point de vue correspondant à chaque nœud :

$$w_{ij} = 1 + \alpha(1 - d(N_{ij}))\ \delta_{ij} \quad (1)$$

avec $\alpha \geq 1$ une constante, $d(N_{ij})$ la distance entre les neurones *i* et *j*, et $\delta_{ij} = 1$ si v_i et v_j appartiennent au même cluster dans C_{SOM} et 0 sinon.

Une fois que les poids sont modifiés selon l'équation 1, une partition, $C_{SOM\text{-}FU}$ est calculée en utilisant le FU. Cette nouvelle partition contient l'ensemble des communautés finales et l'information structurelle. En modifiant les poids du graphe avec l'équation 1, le graphe devient pondéré et les arêtes avec un poids plus grand ont une probabilité plus élevée d'être affectées à la même communauté.

3.3. Visualisation et exploration de communautés

L'objectif de l'algorithme est de placer les nœuds de façon telle que leur proximité indique la similarité existant entre eux. L'algorithme de tracé proposé est divisé en deux étapes : d'abord, placer les nœuds de frontière, puis, dans un deuxième temps, placer les nœuds intérieurs du graphe.

3.3.1. Multi-dimensional scaling

Le *multi-dimensional scaling* (MDS), est une technique pour représenter visuellement des objets en fonction de leur similarité ou de leur dissemblance. La distance dans un espace ρ-dimensionnel, généralement euclidien, respecte la similarité de l'espace d'origine. Cet algorithme permet alors de placer les nœuds selon leur similarité.

On voudrait représenter le graphe de clusters dans un espace bidimensionnel de sorte à voir les interactions entre les différents groupes.

Il existe plusieurs implémentations de l'algorithme MDS qui peuvent être classifiées en trois types (Ingram *et al.,* 2009) : des méthodes classiques cherchant une solution analytique en minimisant une fonction d'effort. Ces méthodes ont une complexité en $O(n^3)$; des méthodes basées sur la distance et l'optimisation non linéaire comme par exemple la méthode de descente de gradient, avec une complexité en $O(L \cdot n^2)$ où L est la dimension cible, et enfin, des méthodes basées sur la simulation des systèmes masse-ressort qui ont une complexité en $O(n^3)$: un processus de complexité $O(n^2)$ exécuté *n* fois.

Dans ce travail, nous utilisons l'algorithme SMACOF *(Scaling by MAjorizing a COmplicated Function),* qui est du deuxième type, et qui converge de façon

monotone en un point stable par réduction d'une fonction de stress (Ingram *et al.,* 2009). La complexité de l'algorithme est en $O(L \cdot n^2)$, où *n* est le nombre d'éléments de l'ensemble.

Pour mesurer l'approximation des distances par rapport aux dissemblances, nous utilisons une fonction de stress. Ainsi, étant donnée une matrice *X* de points dans un espace ρ-dimensionnel, la fonction de stress σ(X) est définie comme suit :

$$\sigma(X) = \sum_{i<j} (d_{ij} - \delta_{ij})^2 \quad (2)$$

où d_{ij} est la distance et δ_{ij} est la dissemblance entre les objets *i* et *j*.

3.3.1.1. Mesures de dissemblance

Dans le cas étudié ici, nous cherchons à représenter les nœuds d'un graphe. Pour utiliser le MDS, nous devons définir la dissemblance δ_{ij} entre nœuds. Une mesure très utilisée (Wasserman et Faust, 1994) est la distance géodésique entre les nœuds : $d(u,v) = \min_p A_{uv}^{[p]}$, où *A* est la matrice d'adjacence et $A^{[p]}$ est la matrice puissance *p*. Cela représente le chemin le plus court de longitude *p* entre les nœuds *u* et *v*.

Un autre type de mesure englobe les métriques basées sur la disparité des voisinages de deux nœuds donnés. Deux mesures de ce type sont la distance de Jaccard et la distance cosinus (Fortunato, 2010).

Soit *N(u)* l'ensemble des voisins du nœud *u*. Alors, étant donnés deux nœuds u, v $\in$ V, la distance de Jaccard d_J est donnée par $d_J(u,v) = 1 - \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$.

Des voisinages similaires indiquent que les nœuds sont aussi similaires.

3.3.1.2. Initialisation des points

Puisque le MDS utilise un algorithme de descente de gradient, les valeurs initiales des points peuvent changer le résultat final. Cela veut dire que le résultat est un minimum local dépendant du point de départ. Nous définissons une matrice $X_{[0]}$ avec les coordonnées initiales des nœuds. Ces coordonnées peuvent être définies de façon aléatoire, mais le dessin obtenu sera alors différent d'une exécution à l'autre. Or nous voulons un dessin identique, pour le même ensemble de données, à chaque exécution, de manière à préserver une stabilité visuelle pour l'utilisateur. Nous utilisons pour ce faire une procédure qui place chaque point sur une circonférence de rayon $r = 1$ et de centre $c = (x_0, y_0)$. Ensuite, la position du nœud *i* est :

$$X_x^{[a]}(i) = \cos \theta_i \quad (3)$$
$$X_y^{[a]}(i) = \sin \theta_i$$

où $\theta_i = 2\pi(i-1)/k$, $1 \leq i \leq k$.

Cette initialisation suit la recommandation de Borg et Groenen (1997) et permet d'avoir toujours la même position initiale pour les nœuds.

3.3.2. Dessin des nœuds de frontière V^+

Les nœuds de frontière sont placés dans un cercle de rayon 0,5. Ce cercle est centré sur l'origine absolue c = (0, 0), qui est utilisée par l'algorithme SMACOF comme centre de référence.

L'algorithme 1 montre l'implémentation pour dessiner les nœuds de frontière. Cet ensemble est pris en entier et contient les nœuds de tous les groupes. L'idée est de montrer les relations entre les communautés grâce à la position de chaque nœud frontière. Comme la dissemblance entre nœuds est calculée selon leur voisinage, les positions proches permettent de voir leur proximité structurelle dans le graphe *G*. Ainsi des nœuds qui sont à l'interface entre les mêmes communautés auront des rôles similaires de médiation par exemple.

Algorithme 1. Algorithme de localisation des nœuds de frontière

```
1:  dessinNœudsFrontière (V+)
2:  {
3:      Δ_V+ ← d_J (V+)
4:      X_V+ ← SMACOF(Δ_V+)
5:      return X_V+
6 : }
```

L'algorithme retourne X_{V+}, l'ensemble des coordonnées pour chaque nœud de frontière. Une fois que les positions sont déterminées, elles sont modifiées de façon à ce que la distance de chaque point à l'origine soit inférieure à 0,5. Cette normalisation permet de bien séparer les zones destinées à chaque type de nœud.

3.3.3. Dessin des nœuds intérieurs V_i^-

V_i^- est le sous-ensemble des nœuds intérieurs à la communauté *i*. Chacun de ces sous-ensembles doit être placé près des nœuds de frontière déjà placés, qui appartiennent à la même communauté. Pour ce faire, nous définissons le centre P_i de la communauté *s* comme :

$$P_i^x = \pm r * \sqrt{(1/(m_i^2+1))} \qquad (4)$$
$$P_i^y = P_i^x * m_i$$

où r = 0,75 est le rayon du cercle au milieu de l'anneau défini par l'espace des nœuds de frontière et la limite de la surface de dessin, soit le cercle $x^2 + y^2 = 1$, et m_i est la pente du vecteur formé par le centre de masse des nœuds frontière de la communauté *i* et l'origine. Donc, $m_i = \frac{\overline{y_i}}{\overline{x_i}}$ où $\overline{x_i}$ et $\overline{y_i}$ sont :

$$\left(\overline{x_i}, \overline{y_i}\right) = \sum_{u \in C_i \cap V^+} \frac{u_x, u_y}{\left|C_i \cap V^+\right|} \qquad (5)$$

où u_x, u_y sont les coordonnées x et y du nœud u.

Ainsi, chaque point $P_i = (P_i^x, P_i^y)$ est le centre de sa communauté et sera utilisé par l'algorithme 2 pour placer les nœuds de chaque communauté.

Algorithme 2. Algorithme de localisation des nœuds intérieurs

1: ***dessinNœudsInterieurs*** (V^-)
2: {
3 : **pour tout** i ∈ [1..k] **faire :**
4: $\Delta_{V_i^-} \leftarrow d_J\left(V_i^-\right)$
5: $X_{V_i^-} \leftarrow$ SMACOF($\Delta_{V_i^-}$)
6 : $X_{V_i^-} \leftarrow$ adjustCentre(Pi)
6: **fin boucle**
7 : **}**

De cette manière les nœuds intérieurs de chaque communauté sont placés selon leur ressemblance de voisinage et face aux nœuds de frontière de sa communauté.

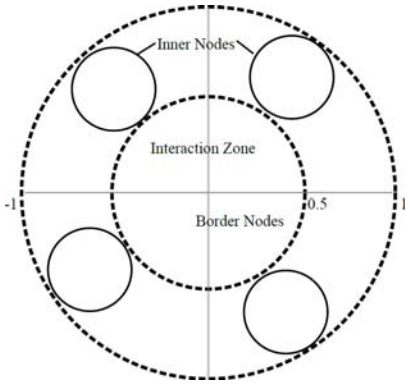

Figure 10. Localisation des éléments du modèle de dessin

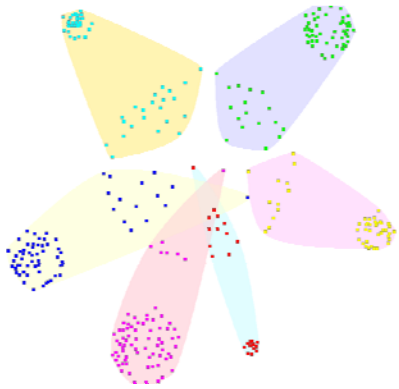

Figure 11. Localisation des nœuds intérieurs selon la position des nœuds de frontière

La Figure 10 présente la disposition visuelle des différents éléments du modèle et la Figure 11 montre un exemple de la localisation finale des nœuds avec l'algorithme décrit.

4. La complexité et le passage à l'échelle

La complexité de ce modèle de détection et de visualisation de communautés est divisée en deux parties : d'abord la complexité liée à la détection de communautés, puis la complexité de l'algorithme de dessin. Dans le premier cas, il faut tenir compte de la combinaison de deux algorithmes exécutés de façon séquentielle dont la somme des complexités est la complexité globale.

Pour le modèle de visualisation, l'algorithme est exécuté k+1 fois : une fois pour placer les nœuds de frontière et *k* fois pour chacun des ensembles des nœuds intérieurs. La complexité globale est alors proportionnelle à la taille de chaque catégorie (les nœuds frontière et intérieurs).

Pour tester la complexité dans les deux cas, nous avons utilisé un graphe composé de 5 389 nœuds et 46 440 arêtes, ainsi qu'un ensemble de 35 caractéristiques pour construire les points de vue.

4.1. Complexité de l'algorithme de détection de communautés

Le premier pas de l'algorithme de détection de communautés est le regroupement sémantique des nœuds en utilisant des points de vue. Ce processus compare chacun des nœuds du graphe avec les $\eta = f \times f$ neurones de la carte de Kohonen ; le processus est répété jusqu'à la fin de l'entraînement du réseau. La complexité est alors délimitée par :

$$T_{SOM}(n) = O(|PoV_{F*}|^2 \times n) \quad (6)$$

où PoV_{F*} est le point de vue choisi. Ensuite le poids de chaque arête doit être modifié. Cette opération a une complexité de :

$$T_{\Omega}(n) = O(m) \quad (7)$$

où *m* est le nombre d'arêtes du graphe. À la fin l'algorithme *Fast Unfolding* est exécuté. Cet algorithme a une complexité de $T_Q = O(n)$ pour des graphes épars, en conséquence, la complexité globale est (avec les équations (6) et (7)) :

$$T_{Gl} = T_{SOM}(n) + T_{\Omega}(n) + T_Q(n) \quad (8)$$

La Figure 12 présente le temps d'exécution pour l'algorithme de détection de communautés en utilisant des points de vue avec différentes tailles. Notez que l'augmentation de temps est quadratique en fonction de la taille du point de vue.

4.2. Complexité du modèle de visualisation de communautés

Le modèle de visualisation est divisé en deux étapes générales : dans la première, les nœuds de frontière sont placés, puis, pendant la deuxième ce sont les nœuds intérieurs de chaque communauté.

L'algorithme SMACOF utilise une approche de descente de gradient pour trouver l'ensemble des coordonnées dont les distances ressemblent aux similarités d'un ensemble de points dans une autre dimension. À cause des multiplications matricielles, cet algorithme a une complexité générale de $T(n) = O(L \cdot n^2)$ où L est la taille de la dimension cible, dans notre cas L = 2. Toutefois, grâce à la division de l'ensemble de nœuds en catégories, le calcul de la complexité change. La complexité pour l'ensemble des nœuds de frontière est :

$$T_\beta(n) = O(2 \cdot |V^+|^2) \tag{9}$$

et la complexité pour les ensembles des nœuds intérieurs est :

$$T_i(n) = O\left(\sum_{i=1}^{k} 2 \cdot \left| V^- \right|^2 \right) \tag{10}$$

La taille espérée de chaque ensemble de nœuds intérieurs est $(n-|V^+|) / k$, donc il est possible de réécrire l'équation (10) sous la forme :

$$T_i(n) = O\left(\frac{2 \cdot \left(n - \left| V^+ \right| \right)^2}{k^2} \right) \tag{11}$$

où *k* est le nombre de communautés dans le graphe. Ainsi, la complexité générale est :

$$\hat{T}(n) = O\left(\frac{2 \cdot \left(n - \left| V^+ \right| \right)^2}{k^2} + 2 \cdot \left| V^+ \right|^2 \right) \tag{12}$$

Notons qu'en général $\hat{T}(n) < T(n)$, en conséquence le pire cas se produit quand tous les nœuds appartiennent à l'ensemble des nœuds de frontière, où la complexité devient en $O(2 \cdot n^2)$.

Dans les réseaux sociaux ce cas particulier n'est pas très probable d'après la structure de ce type de réseaux. Selon Newman and Girvan (2004) ces graphes ont des structures de groupes avec un nombre plus élevé de liens dans les communautés qu'en dehors. La Figure 13 présente le temps d'exécution de l'algorithme en changeant le nombre de nœuds de frontière : le temps explose quand la majorité du calcul concerne ces nœuds.

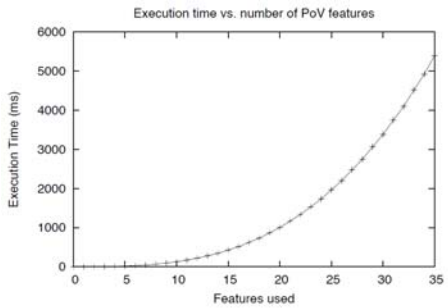

Figure 12. Temps d'exécution pour l'algorithme de détection de communautés

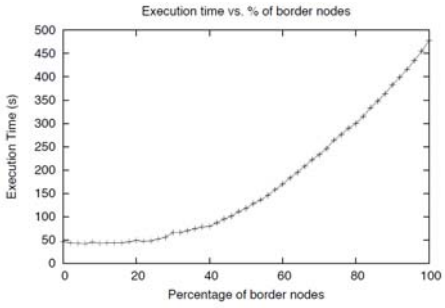

Figure 13. Temps d'exécution pour l'algorithme de dessin de communautés

5. Expériences et résultats

Nous proposons une série d'expériences afin de tester l'algorithme de détection et de visualisation de communautés et pour effectuer quelques analyses à partir de ces résultats. Dans ces expériences, nous examinons le comportement des partitions quand les types d'information sont intégrés, d'après une perspective de composition des groupes et d'une perspective visuelle.

5.1. Configuration des expériences

Pour exécuter les expériences nous avons utilisé un réseau social extrait de Facebook. Ce jeu de donnés représente un réseau social personnel récolté avec NameGenWeb (Hogan, 2011) et inclut des personnes proches de son propriétaire : famille, amis, collègues et anciens collègues. Au total le réseau a 334 nœuds et 5 394 arêtes.

Avec ce graphe nous avons défini deux points de vue. Un sans caractéristique (PoV_{NULL}), qui représente uniquement les résultats du modèle sur le graphe, et un autre point de vue (PoV_{COMP}) qui décrit les compétences de chaque acteur. Le Tableau 1 résume la composition de chaque point de vue utilisé pour les expériences.

5.2. Description des résultats de la détection de communautés

Pour mesurer la qualité des partitions nous utilisons deux mesures : la *modularité* Q et la *distance sémantique moyenne* $\overline{D_c}$. La première, $Q \in [-1,1]$, proposée par (Newman et Girvan, 2004), compare la fraction des arêtes dans les groupes versus le nombre d'arêtes connectant les groupes. Cette mesure a été conçue pour des graphes

épars avec une structure de communautés (Blondel *et al.,* 2008) ; pour d'autres types de graphes, cette mesure peut donner des résultats inattendus.

La deuxième mesure, $\overline{D_c}$, calcule la distance moyenne entre tous les nœuds de chaque groupe. Cette distance est normalisée pour que tous les résultats soient dans l'intervalle [0, 1] : 0 signifie que les nœuds sont complètement similaires et 1 qu'ils sont tout à fait différents.

Le modèle est utilisé avec le graphe et les points de vue présentés dans le Tableau 1, les résultats sont résumés dans le Tableau 2.

Tableau 1. Résumé des points de vue utilisés pendant les expériences

PoV	Catégories	Explication
NULL	Aucune	Le PoV_{NULL} ne contient pas de catégorie car les algorithmes sont appliqués sur la structure du graphe uniquement.
COMP	1. Math & Science 2. Business Administration 3. Law 4. Social Sciences 5. Software engineering 6. Other fields 7. Arts	Ces catégories ont été définies selon les champs d'action de chaque acteur.

Tableau 2. Résumé des résultats du processus de détection de communautés

PoV	Groupes	Modularité	Distance Sémantique Moyenne ± écart type
NULL	6	0,7728	Par rapport à PoV_{COMP} : 0,4689 (±0,0201)
COMP	10	0,8138	0,2820 (±0,0431)

Le point de vue PoV_{NULL} a la valeur optimale de modularité car l'algorithme prend en compte seulement la structure du graphe ; cette valeur devient alors notre valeur de référence. La distance sémantique moyenne du PoV_{NULL} symbolise l'état de la partition structurelle en termes de l'autre point de vue.

Pour le point de vue PoV_{COMP}, la valeur de modularité est supérieure à celle du PoV_{NULL}. Cela veut dire que le changement des poids du graphe a eu un effet sur l'exécution de l'algorithme de détection de communautés, en faisant que les nœuds les plus similaires ont été affectés au même groupe. Ce fait peut être corroboré par la valeur de $\overline{D_c}$ (dernière colonne) inférieure à celle du PoV_{NULL}.

La distribution des catégories dans chaque partition est représentée dans les figures 14 et 15. Dans le premier cas, les groupes sont identifiés à partir de l'information structurelle. En conséquence, chaque partition contient des nœuds provenant de presque toutes les catégories, à l'exception des groupes 2 et 3.

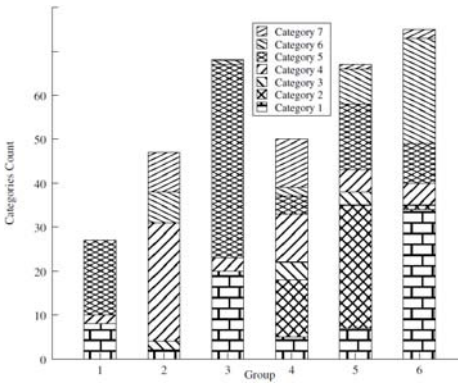

Figure 14. Distribution des catégories dans la partition PoV_{NULL}

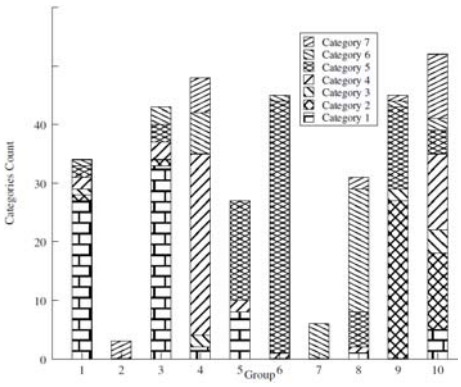

Figure 15. Distribution des catégories dans la partition PoV_{COMP}

Le deuxième cas inclut l'information sémantique dans le processus de détection de communautés, ce qui fait que chaque groupe a une catégorie dominante et un contenu moins variable.

La caractéristique principale du modèle est l'intégration de l'information utilisée pour identifier les partitions avec des critères structuraux et sémantiques. Néanmoins, l'inclusion d'information sémantique pour modifier la structure du graphe impose quelques restrictions sur la qualité de la structure de communautés du graphe : ni la valeur de modularité, ni la valeur de distance ne sont optimales, mais il s'agit d'un compromis entre les deux, comme cela a été présenté dans Cruz *et al.,* (2011b).

En utilisant la partition PoV_{COMP} il est possible d'extraire de nouvelles informations qui n'étaient pas évidentes dans PoV_{NULL}. Par exemple, le groupe 2 contient des personnes qui étaient dans le groupe 2 de PoV_{NULL} (famille), mais leur compétence est très spécifique, donc ils forment un nouveau groupe dans PoV_{COMP}.

Le groupe 7 de PoV_{COMP} contient des anciens collègues d'un projet de recherche. Ces personnes partagent la même compétence et sont aussi fortement connectés.

Ainsi, le modèle est capable de diviser des partitions structurelles afin de mieux explorer chaque groupe depuis une optique sémantique.

5.3. Description des résultats de la visualisation

Pour analyser les résultats de la visualisation des partitions, nous utilisons une catégorisation des rôles des nœuds proposée par Cross et Parker (2004) :

– nœuds isolés : des nœuds qui sont déconnectés du reste du réseau ou qui sont connectés seulement avec un autre nœud ;

– nœuds centraux : des nœuds avec un nombre très important de connexions, importants en termes de diffusion d'information et de concentration de liens ;

– nœuds d'ouverture : nœuds qui permettent la communication des nœuds d'un groupe vers l'extérieur ou de l'extérieur vers l'intérieur du groupe ;

– nœuds d'intermédiation : nœuds qui partagent des connexions avec différentes communautés ou qui font office de liens entre des communautés différentes.

Cette catégorisation nous permet d'analyser par la suite chaque point de vue d'une perspective visuelle.

5.3.1. Nœuds isolés

Ces nœuds ont un degré 0 ou 1, et selon la division des nœuds du modèle de visualisation, ils sont des nœuds intérieurs. La figure 16 présente les nœuds isolés dans la partition PoV_{NULL} et la Figure 17 présente les nœuds isolés dans la partition PoV_{COMP}.

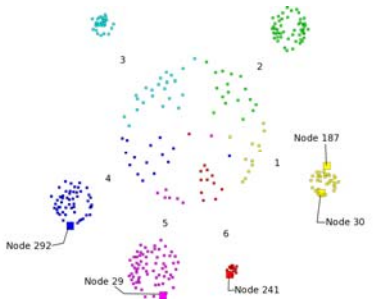

Figure 16. Nœuds isolés de la partition NULL

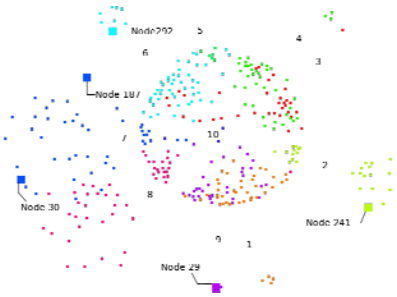

Figure 17. Nœuds isolés de la partition COMP

La condition d'isolement est liée au degré du nœud, par conséquent en changeant le point de vue on ne change pas les nœuds signalés comme isolés. Le seul changement est la position dans le dessin, qui change grâce au changement de la configuration des partitions d'un point de vue à l'autre. Par exemple, les nœuds 187 et 30 dans la Figure 16 sont localisés dans le groupe 1 (famille et amis), mais quand le point de vue est changé, ils sont déplacés vers le groupe 7 (figure 17) où la plupart des gens appartiennent à la catégorie sémantique 4.

5.3.2. Nœuds centraux

Ce type de nœuds concentre une partie importante des connexions du réseau : ils centralisent le passage de l'information à travers le réseau. Pour les identifier nous choisissons ceux qui ont un degré de connectivité supérieur à celui de la majorité des nœuds : c'est-à-dire dont le degré est supérieur à $[\mu+3\sigma]$, où μ est le degré moyen et σ est l'écart type.

La figure 18 présente la localisation des nœuds centraux pour le point de vue NULL et la figure 19 montre la localisation de ces nœuds dans la partition du point de vue COMP.

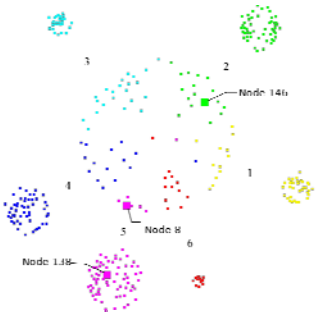

Figure 18. Nœuds centraux dans la partition NULL

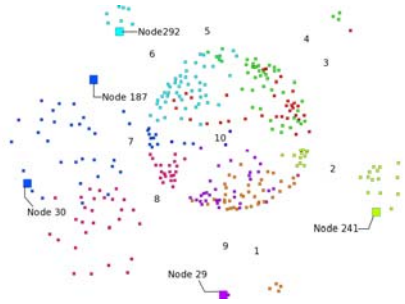

Figure 19. Nœuds centraux dans la partition COMP

La caractéristique des nœuds centraux est leur grand nombre de connexions, donc le fait de changer le point de vue ne change pas cet ensemble de nœuds, par contre, il peut changer leur type. Par exemple, le nœud 138 dans le PoV_{NULL} était un nœud intérieur, mais après l'utilisation du PoV_{COMP} il est devenu nœud de frontière.

En plus, l'utilisation du PoV_{COMP} a changé la configuration des communautés, notamment l'ancienne communauté 5 (PoV_{NULL}), composée de gens rencontrés pendant des études de doctorat, a été divisée en deux, une d'étudiants et une autre de chercheurs : cela montre comment l'utilisation d'un point de vue change le niveau de résolution de chaque communauté.

D'un autre côté, le nœud 146 reste presque dans la même position à chaque point de vue : ce nœud appartient à une communauté où la plupart des personnes appartiennent à la même catégorie dans les deux PoV.

5.3.3. Nœuds d'ouverture

Ces nœuds connectent des nœuds intérieurs avec la zone d'interaction : ils permettent aux nœuds intérieurs de se connecter avec le monde « extérieur ». Ces nœuds sont placés au centre du dessin et sont classifiés comme nœuds de frontière. La figure 20 présente les nœuds d'ouverture correspondant à la partition NULL et la figure 21 présente les nœuds d'ouverture dans la partition COMP.

Le premier effet du changement de point de vue est la réduction du nombre de nœuds d'ouverture. Pour la partition NULL, le nombre de ces nœuds est de 53, alors que pour la partition COMP le nombre est de 33. Cette différence est due au changement des nœuds intérieurs en des nœuds de frontière, ce que reflète la réduction de la modularité lorsque la structure du graphe est changée.

Un exemple de ce changement de typologie est le nœud 185 : dans la figure 20 il est de type intérieur, alors que dans la figure 21 il est placé dans la zone d'interaction comme un nœud de frontière.

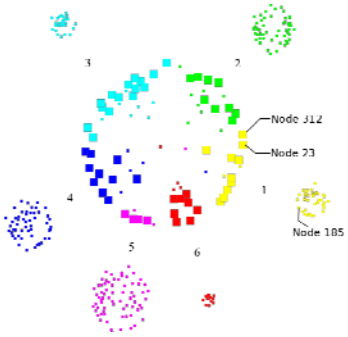

Figure 20. Nœuds d'ouverture dans la partition NULL

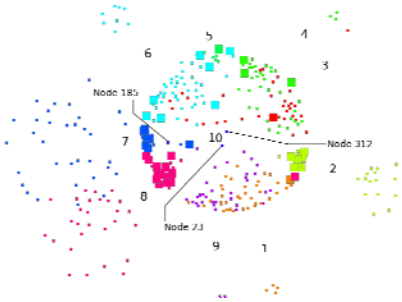

Figure 21. Nœuds d'ouverture dans la partition COMP

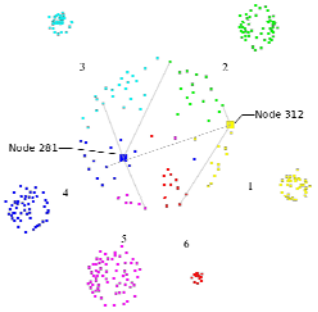

Figure 22. Nœuds d'intermédiation dans la partition NULL

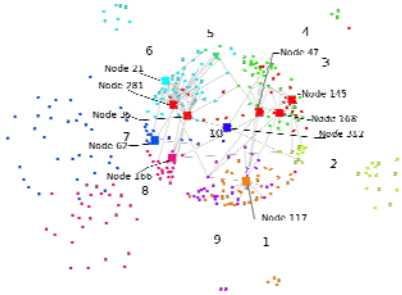

Figure 23. Nœuds d'intermédiation dans la partition COMP

5.3.4. Nœuds d'intermédiation

Des nœuds possédant ce rôle connectent différentes communautés dans la zone d'interaction, c'est-à-dire qu'ils établissent un pont entre deux communautés ou plus. La figure 22 présente la localisation des nœuds d'intermédiation pour la partition NULL, la figure 23 présente les nœuds d'intermédiation pour la partition COMP.

La partition NULL contient uniquement deux nœuds d'intermédiation, qui connectent les différentes partitions du graphe. Dans la partition COMP il existe 10 nœuds de ce type dont deux sont les mêmes que dans la partition NULL et 8 sont nouveaux.

La différence fondamentale entre les nœuds d'intermédiation et les nœuds d'ouverture est ce qu'ils connectent : les premiers permettent l'interaction entre communautés, alors que les deuxièmes permettent aux nœuds intérieurs à leur propre communauté de se connecter avec la zone d'interaction. Notez que ces rôles ne sont pas exclusifs et qu'un nœud d'ouverture peut être aussi d'intermédiation.

6. Conclusion et perspectives

L'information contenue dans un réseau social peut être divisée en trois catégories : structurelle, de composition ou sémantique et d'affiliation. En général la plupart des algorithmes de détection et de visualisation des communautés n'utilisent que l'information structurelle pour identifier les groupes.

Nous avons présenté un modèle de détection et de visualisation de communautés qui intègre les variables structurelles et de composition pour générer un nouveau type de variable d'affiliation qui représente des partitions structurelles et sémantiques en groupes. Ainsi, l'information structurelle est représentée par le graphe social et l'information sémantique par des caractéristiques liées à chaque nœud.

Le modèle de détection, crée premièrement une partition avec l'information sémantique. Cette partition est définie par des groupes sémantiquement proches, c'est-à-dire similaires selon leurs caractéristiques. Ensuite, avec cette partition sémantique, les poids du graphe sont changés de façon telle que les arêtes connectant deux nœuds dans le même groupe sémantique aient une probabilité plus élevée d'être affectés à la même communauté. Le changement des poids permet d'intégrer l'information sémantique à la structure du graphe.

Ensuite, le nouveau modèle de visualisation de graphes de communautés divise l'ensemble des nœuds du graphe selon leur type de connectivité : des nœuds avec des connexions depuis/vers d'autres groupes, appelés nœuds de frontière et des nœuds ne comportant des arêtes que dans leur propre communauté. Avec cette division il est possible de visualiser les interactions entre communautés et aussi d'identifier, visuellement, les nœuds ayant des rôles importants pour la communication entre groupes.

La position des nœuds est déterminée par l'algorithme de *multi dimensional scaling*, qui fait une projection des similarités d'un espace ρ-dimensionnel dans un espace $\mathbb{R}^2$ avec des distances euclidiennes. La similarité est calculée à partir du voisinage de chaque nœud : deux nœuds sont similaires si leur voisinage est similaire, ainsi, deux nœuds similaires vont être proches dans le plan $\mathbb{R}^2$.

Pour tester le modèle nous avons utilisé un graphe extrait de Facebook et un point de vue créé à partir des compétences des personnes qui composent ce graphe. Le point de vue décrit donc chaque nœud en fonction d'une des sept catégories liées à un champ de connaissances.

Les résultats de la détection de communautés montrent que l'influence du point de vue réduit la distance sémantique des groupes au détriment de la modularité : il

existe un compromis entre les deux mesures de qualité à cause de la nature contradictoire de chacune de ces mesures.

La visualisation des graphes de communautés permet, d'un côté, d'identifier et de mettre en évidence les relations et interactions entre les communautés autant que les différents rôles d'interaction. D'un autre côté, le modèle visuel montre la séparation des groupes, comme un effet de la baisse de la modularité, mais aussi de la différenciation catégorielle de chaque communauté.

En ce qui concerne les temps d'exécution, le modèle a un comportement quadratique en fonction du nombre de caractéristiques pour la détection de communautés et quadratique également en fonction du nombre de nœuds de frontière pour la visualisation.

Les perspectives de ces travaux consistent à étudier le passage d'un point de vue à un autre de façon à pouvoir comparer les caractéristiques visuelles et de composition. Nous prévoyons également de travailler sur le développement d'un outil de navigation et d'exploration qui permettrait d'analyser l'influence des points de vue dans la visualisation.

Bibliographie

Blondel V.D., Guillaume J.-L., Lambiotte R., Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, No. 10, pp. P10008 (12pp). http://stacks.iop.org/1742-5468/2008/P10008

Borg I., Groenen P. (1997). *Modern multidimensional scaling : theory and applications* (Springer, Ed.). New York, N.Y., Springer.

Bourqui R., Auber D., Mary P. (2007, July). How to draw clustered-weighted graphs using a multilevel force-directed graph drawing algorithm. *Information visualization, 2007. iv '07. 11th international conference*, p. 757-764.

Brockenauer R., Cornelsen S. (2001). Drawing clusters and hierarchies. In M. Kaufmann, D. Wagner (Eds.), *Drawing graphs,* vol. 2025, p. 193-227. Springer Berlin/Heidelberg. http://dx.doi.org/10.1007/3-540-44969-8-8

Cross R., Parker A. (2004). *The hidden power of social networks: Understanding how work really gets done in organizations* (H.B.S. Press, Ed.). Harvard Business School Press.

Cruz J. D., Bothorel C., Poulet F. (2011, october). Entropy based community detection in augmented social networks. *Computational aspects of social networks (cason), 2011 international conference on*, p. 163-168.

Di Giacomo E., Didimo W., Grilli L., Liotta G. (2007, March-April). Graph visualization techniques for web clustering engines. *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, n° 2, p. 294-304.

Du N., Wu B., Pei X., Wang B., Xu L. (2007). Community detection in large-scale social networks. *Webkdd/sna-kdd'07: Proceedings of the 9th webkdd and 1st sna-kdd 2007 workshop on web mining and social network analysis*, p. 16-25. New York, USA, ACM.

Eades P., Feng Q.-W. (1997). Multilevel visualization of clustered graphs. In S. North (Ed.), *Graph drawing*, vol. 1190, p. 101-112. Springer Berlin/Heidelberg. http://dx.doi.org/10.1007/3-540-62495-3-41

Fortunato S. (2010). Community detection in graphs. *Physics Reports*, vol. 486, n° 3-5, p. 75-174. http://www.sciencedirect.com/science/article/B6TVP-4XPYXF1-1/2/99061fac6435db4343b2374d26e64ac1

Gaetler M. (2005). *Network analysis: Methodological foundations.* U. Brandes, T. Erlebach (Eds.), p. 178-215. Springer Berlin/Heidelberg.

Hogan B. (2011). *Namegenweb*. Facebook Application. http://apps.facebook.com/namegenweb/

Hughes G.F. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, vol. 14, n° 1, p. 55-63. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1054102

Ingram S., Munzner T., Olano M. (2009). Glimmer: Multilevel mds on the gpu. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 15, pp. 249-261.

Kohonen T. (1997). *Self-organizing maps*. Springer.

Lipczak M., Milios E. (2009). Agglomerative genetic algorithm for clustering in social networks. *Gecco'09: Proceedings of the 11th annual conference on genetic and evolutionary computation*, p. 1243-1250. New York, NY, USA, ACM.

Newman M.E.J., Girvan M. (2004, Feb.). Finding and evaluating community structure in networks. *Physical Review. E, Statistical Nonliner and Soft Matter Physics*, vol. 69, n° 2, p. 026113.

Noack A. (2003). An energy model for visual graph clustering. *Proceedings of the 11th int. symposium on graph drawing (gd 2003), lncs 2912*, p. 425-436. Springer-Verlag.

Pizzuti C. (2009). Overlapped community detection in complex networks. *Gecco'09: Proceedings of the 11th annual conference on genetic and evolutionary computation*, p. 859-866. New York, NY, USA, ACM.

Rousseeuw P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. of Computational and Applied Mathematics*, vol. 20, n° 1, p. 53-65.

Santamaría R., Therón R. (2008). Overlapping clustered graphs: Co-authorship networks visualization. A. Butz, B. Fisher, A. Krüger, P. Olivier, M. Christie (Eds.), *Smart graphics*, vol. 5166, p. 190-199. Springer Berlin/Heidelberg. http://dx.doi.org/10.1007/978-3-540-85412-8-17

Shi J., Malik J. (2000). Normalized cuts and image segmentation. *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 22, n° 8, p. 888-905.

Tamassia R. (1987, June). On embedding a graph in the grid with the minimum number of bends. *SIAM J. Comput*., vol. 16, p. 421-444. http://dx.doi.org/10.1137/0216030

Wasserman S., Faust K. (1994). *Social network analysis: Methods and applications* n° 8. Cambridge University Press.

Zhou Y., Cheng H., Yu J. X. (2009, August). Graph clustering based on structural/attribute similarities. *Proc. VLDB Endow.,* vol. 2, p. 718-729. http://portal.acm.org/citation.cfm?id=1687627.1687709